DOE/NV/25946--066

Title: Error estimates derived from the data for least-squares spline fitting
Author: Jerome Blair
Affiliation: National Security Technologies, LLC
Keywords: data smoothing, multiresolution analysis, spline functions, estimation

Introduction

This is the first in a series of papers on a particular class of practical methods for extracting an accurate estimate of a signal from noisy measurements. The problem, in the simplest form that will be considered, is that a signal $s(t)$ is measured at uniformly spaced discrete times t_i for $i = 1$ to N. The measurements have random noise with known statistics. Throughout this paper it will assumed that the measurement noise is white. However, for a few years, the author been successfully using these methods for problems in which the noise is not white and not even stationary, and the sampling very nonuniform. This problem was first systematically studied in its modern form in [1]-[3], though closely related problems were studied by Gauss [4] as far back as 1804. The measured signal is represented as $y = s + e$, where s is the true signal and e is the vector of measurement errors. We estimate the signal with $\hat{s}$, where

$$\hat{s} = Py = Ps + Pe\,, \tag{1}$$

where P is a linear operator that is applied to the data to give an accurate estimate of the signal. The operator, P, is designed to smooth, or filter, the data to reduce the noise while not distorting the signal too much. The error in the recovered signal is given by e_s with

$$e_s = s - \hat{s} = s - Ps - Pe = (I - P)s - Pe\,, \tag{2}$$

where I is the identity operator. In Equation (2) there are two sources of error, one resulting from the term $(I\text{-}P)s$ and one resulting for the term Pe. In this paper the first term is called the F-error (which could equally mean fitting error or filtering error). It is the error the smoothing operation introduces in the absence of measurement errors. The second term is called the R-error, which is the error in the reconstructed signal caused by the measurement errors. Since it is assumed that there is a known statistical distribution for e, the statistical distribution for Pe, the R-error, can be calculated. This paper deals entirely with estimating the F-error.

In [1]-[3] and hundreds of papers and textbooks written since then, a statistical distribution is assumed known for s, thus allowing the computation of a statistical distribution for $(I\text{-}P)s$, the F-error. In this situation it is possible, for any fixed P, to calculate the statistical distribution for e_s and select an optimum P using the minimum mean squared error (MMSE) criterion. The solution to this problem is well known (*c.f.* Chapter 12 of [5]).

It will be shown how to estimate statistical properties of the F-error without any *a priori* statistical knowledge about the signal. Of course, some knowledge about the signal must be assumed. It is assumed that the sampling rate is more than adequate to represent the

signal. Implicit in some of our calculations is that the sampling rate is a factor of five more than the minimum necessary. This assumption is relevant to practical problems, because in recent years the sampling rate and bandwidth of digital oscilloscopes has been increasing rapidly, but the noise level has remained constant or deteriorated. This makes the situation of high sampling rate and high noise level one of importance. It is also assumed that the unknown signal has four derivatives but no assumptions are made about the magnitudes of the derivatives.

The smoothing operators used are based on cubic spline functions. Let the interval over which the signal is measured be $T_1 \le t \le T_2$, and let K be a sequence of time values, t_k, for $k = 1$ to $n < N$ satisfying $t_1 = T_1$, $t_{k+1} > t_k$ and $t_n = T_2$. A *cubic spline* with knots, K, is a function defined on the interval $[T_1, T_2]$ that is a polynomial of degree three on each sub-interval of the form $[t_k, t_{k+1}]$ and that has two continuous derivatives throughout the interval $[T_1, T_2]$. The symbol, S_K, denotes the vector space of cubic spline functions with knot sequence K. Because of the continuity requirement on the second derivatives, the dimension of S_K is n+2. These functions and many algorithms for dealing with them are described in [6]. The algorithms in [6] are given in FORTRAN. The author used the MATLAB implementation of these algorithms [7].

Proposed Approach

The estimate, $\hat{s}$, for the signal is the least squares fit to the data by a cubic spline with a selected knot sequence K. Precisely,

$$\hat{s} \in S_K \text{ and minimizes } \sum_{i=1}^{N} (\hat{s}(t_i) - y_i)^2 . \tag{3}$$

The solution to this problem depends linearly on the data, and is written as

$$\hat{s} = P_K y . \tag{4}$$

Of particular importance is how close the knots are to their neighbors. This is measured with the quantities

$$\Delta_k = t_{k+1} - t_k \text{ for } 1 \le k \le n-1 \text{ and } t_k \in K . \tag{5}$$

The quantity Δ_k is called the *mesh size* of the k^{th} interval. The knot sequences are restricted to those for which Δ does not vary too rapidly, specifically it is required that

$$1/2 \le \Delta_{k+1} / \Delta_k \le 2 \tag{6}$$

The operation (4) is a time-varying low-pass filter with bandwidth of (see [8] and [9])

$$BW \cong \frac{1}{2\Delta_k} \text{ for } t \text{ near the knot } t_k . \tag{7}$$

The optimal filtering will have the mesh size smaller where larger bandwidth is required to represent the signal and larger where smaller bandwidth is required. For a given knot sequence, the R-error can be directly numerically evaluated from the statistical

distribution of the errors. In this paper, a method for estimating the *F*-error is given and analyzed. In particular, for any time, *t*, it is estimated with

$$e_t = \sqrt{\mathrm{E}[(\hat{s}(t) - s(t))^2]} \quad \text{and}$$
$$e'_t = \sqrt{\mathrm{E}[(\hat{s}'(t) - s'(t))^2]} \quad , \tag{8}$$

using only the data, without assuming statistics for the unknown signal. Here E is the expectation operator and the primes denote the derivative with respect to time. The approach is as follows:

1. Construct the alternate knot sequence K_a with knots at T_1 and T_2 and halfway between the knots in K. Note that K_a has one more knot than K, and that, except near the two endpoints of the data, the local mesh size is the same for the two knot sequences.
2. Calculate the alternate signal estimate, $\hat{s}_a(t)$, by applying (4) with the alternate knot sequence.
3. Calculate $\lambda(t) = \max\{|\hat{s}_a(t') - \hat{s}(t')| : \text{for } t - \Delta_k / 2 \le t' \le t + \Delta_k / 2\}$, where Δ_k is the largest of the two mesh sizes (from K and K_a) for the interval between two knots that contains t.
4. Let $e_t = 0.5\lambda(t)$.

The steps above give e_t. To obtain e'_t, replace the function values with the derivatives in Step 3 and change the constant in Step 4 from 0.5 to 0.4.

The paper will describe the intuitive basis for this approach, which is due to the fact that the errors for the two knot sequences have the opposite signs. Given here the results of Monte Carlo simulations that show how well the approach works.

Results

The method given in the previous section for estimating the filtering error was tested on over100 000 signals using Monte Carlo simulations. The simulations were done with uniform knot spacing. It can easily be shown that the results, as presented here, are independent of the knot spacing. The filtering approach has no error for signals that are polynomials of degree three. Thus, by the Peano kernel theorem ([10] page 43 and [11] page 25), the approximation error depends only on the fourth derivative of the signal and depends linearly on it. The simulations were performed for signals whose fourth derivative was a Gaussian noise process with *rms* value of one and a power spectral density of the form

$$S(\omega) = \frac{\tau}{\pi} \frac{1}{1 + (\omega\tau)^2} . \tag{9}$$

The corresponding autocorrelation function is

$$R(t) = \exp\left(-\frac{|t|}{\tau}\right). \tag{10}$$

The value of the correlation time, τ, was varied between one-tenth of the mesh size and ten times the mesh size.

For each sample function generated, the filtering operation given by (4) and the errors, $e(t) = \hat{s}(t) - s(t)$ and $e'(t) = \hat{s}'(t) - s'(t)$ were calculated for t at the knots and t halfway between the knots. The error estimates, given by our proposed procedure, were also calculated at each of these times. The ratio of the actual error to the estimated standard deviation was saved in a histogram (one histogram for each value of τ.) Ideally, the *rms* value of each of these histograms would be one. The actual values are shown in Figure 1 for signal errors and in Figure 2 for derivative errors.

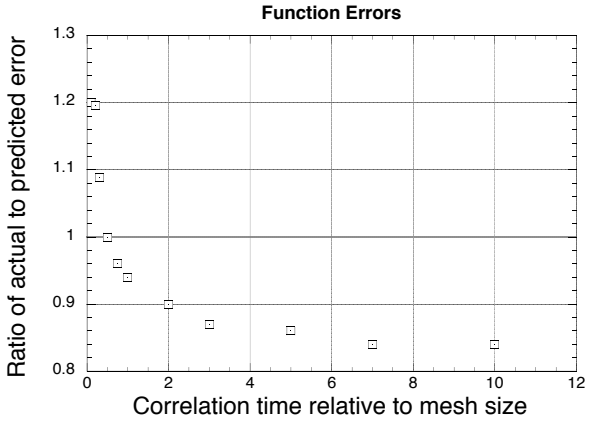

Figure 1

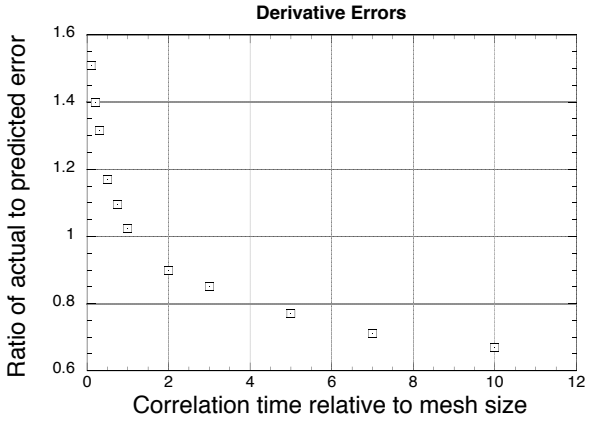

Figure 2

The constants in Step 4 of the estimation procedure were chosen to make the plots more or less symmetric about unity, and they apply to correlation times of one-half to three-quarters of the mesh size. Slightly smaller or larger constants could be used depending on the particular situation. The paper will contain examples using known signals and use

large changes in mesh size over the measurement interval. These will demonstrate the applicability of the method in circumstances not covered by the Monte Carlo simulations.

The study of least squares fitting with splines to reduce noise in measured data appears in Chapter XIV of [6] and in [9]. The idea of varying the knot density to match the local smoothness of the function being approximated is studied in Chapter XII of [6]. However, these studies use *a priori* knowledge about the unknown function rather than the data itself. The determining of good knot sequences based solely on the data was studied in [12] and [13]. However, none of the previously published work produces error estimates for the *F*-error.

References

[1] Weiner, N. *Time Series*, M. I. T. Press, 1949.

[2] Kolmogorov, A, "Interpolation und extrapolation von stationären zufälligen folgen, "*Bulletin de l'académie des sciences de U.R.S.S.*, Ser. Math., pp. 3-14, 1941.

[3] Kosulajeff, P., "Sur les problèmes d'interpolation et d'extraplation des suites stationnaires", *Comptes rendus de l'académie des sciences des U.R.S.S.*, vol. 30, pp. 13-17, 1941.

[4] Gauss, K., *Theory of the Combination of Observations Least Subject to Errors,* SIAM, 1995.

[5] Kay, S., *Fundamentals of Statistical Signal Processing – Estimation Theory*, Prentice Hall, 1993.

[6] de Boor, C., *A Practical Guide to Splines – Revised Edition*, Springer, 2001.

[7] *Spline Toolbox*, The Mathworks.

[8] Unser, M., A. Aldroubi and M. Eden , "Polynomial spline signal approximations: filter design and asymptotic equivalence with Shannon's sampling theorem," *IEEE Trans. Info. Theory*, vol. 38, pp. 95-103, Feb. 1992.

[9] Unser, M., A. Aldroubi, and M. Eden, "B-spline signal processing: part I – theory,"*IEEE Trans. Signal Processing*, vol. 41, pp. 821-832, Feb. 1993.

[10] Ralston, A. and P. Rabinowitz, *First Course in Numerical Analysis,* McGraw-Hill, 1978.

[11] Sard, A., *Linear Approximation*, American Mathematical Society, 1963.

[12] He, X., L. Shen and Z. Shen,"A data-adaptive knot selection scheme for fitting splines," *Signal Proc. Letters*, vol 8, No. 5, pp. 137-139, May 2001.

[13] Ainsleigh, P. and C. Chui, "Simultaneous wavelet and spline smoothing of noisy data," *Acoustics, Speech and Signal Processing, 1993 IEEE International Conference on*, vol. 3, pp. 197-200, Apr. 1993.